Collective Communication on the Quadrics Network (QsNET)

Fabrizio Petrini
fabrizio@lanl.gov

http://www.c3.lanl.gov/~fabrizio

Performance and Architecture Laboratory

CCS-3 Modeling, Algorithms, and Informatics Group

Los Alamos National Laboratory

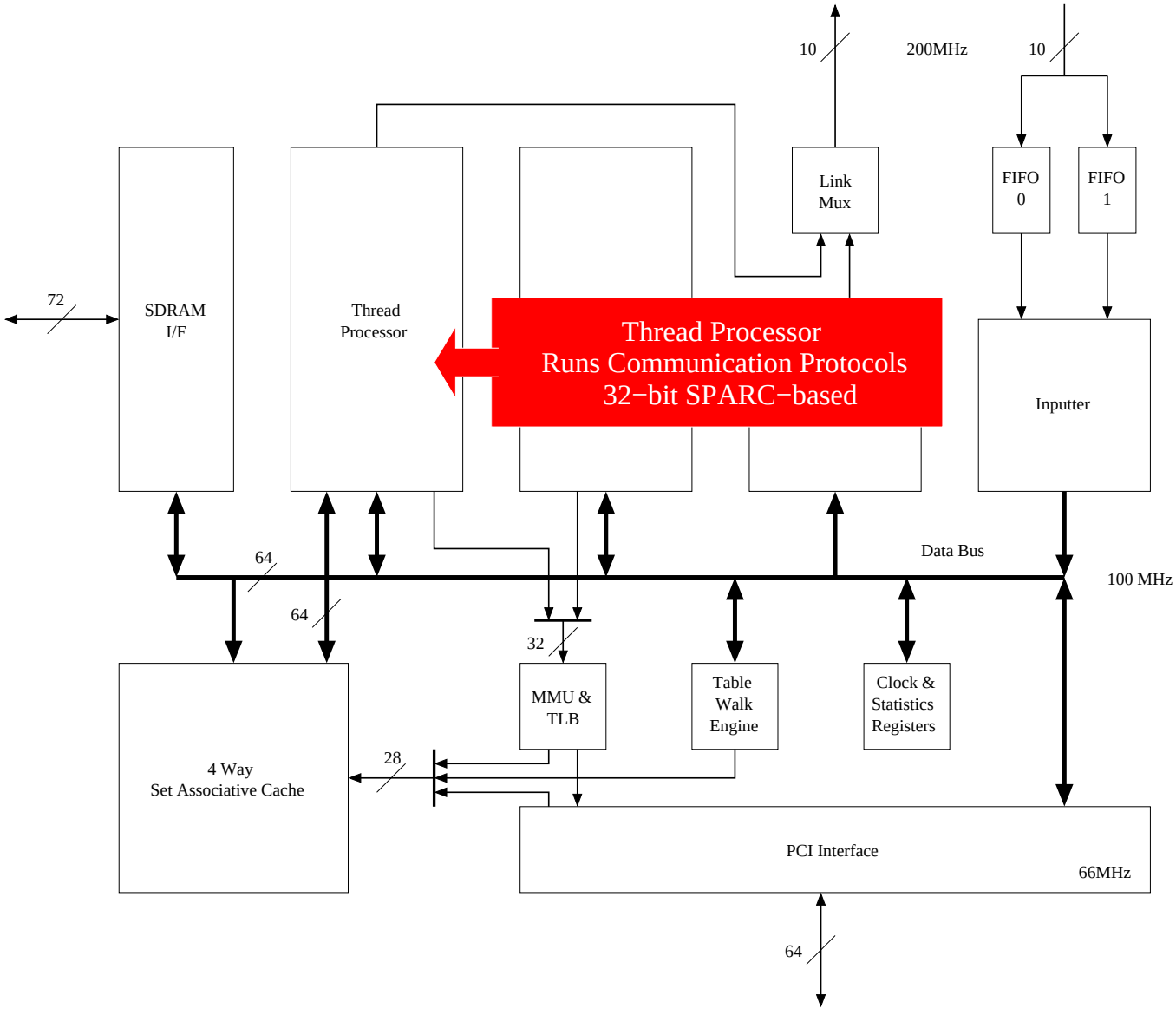
Outline

- Quick overview of the Quadrics network (QsNET)
- Network-based algorithms to perform collective communication
- Hardware support for collective communication
- Performance and scalability results of three collective communication operations (barrier, broadcast and hot spot) on a 1024 node segment of the Q machine
- Ongoing work on *allreduce*, fully implemented in the network, through emulated floating point in the network interface card

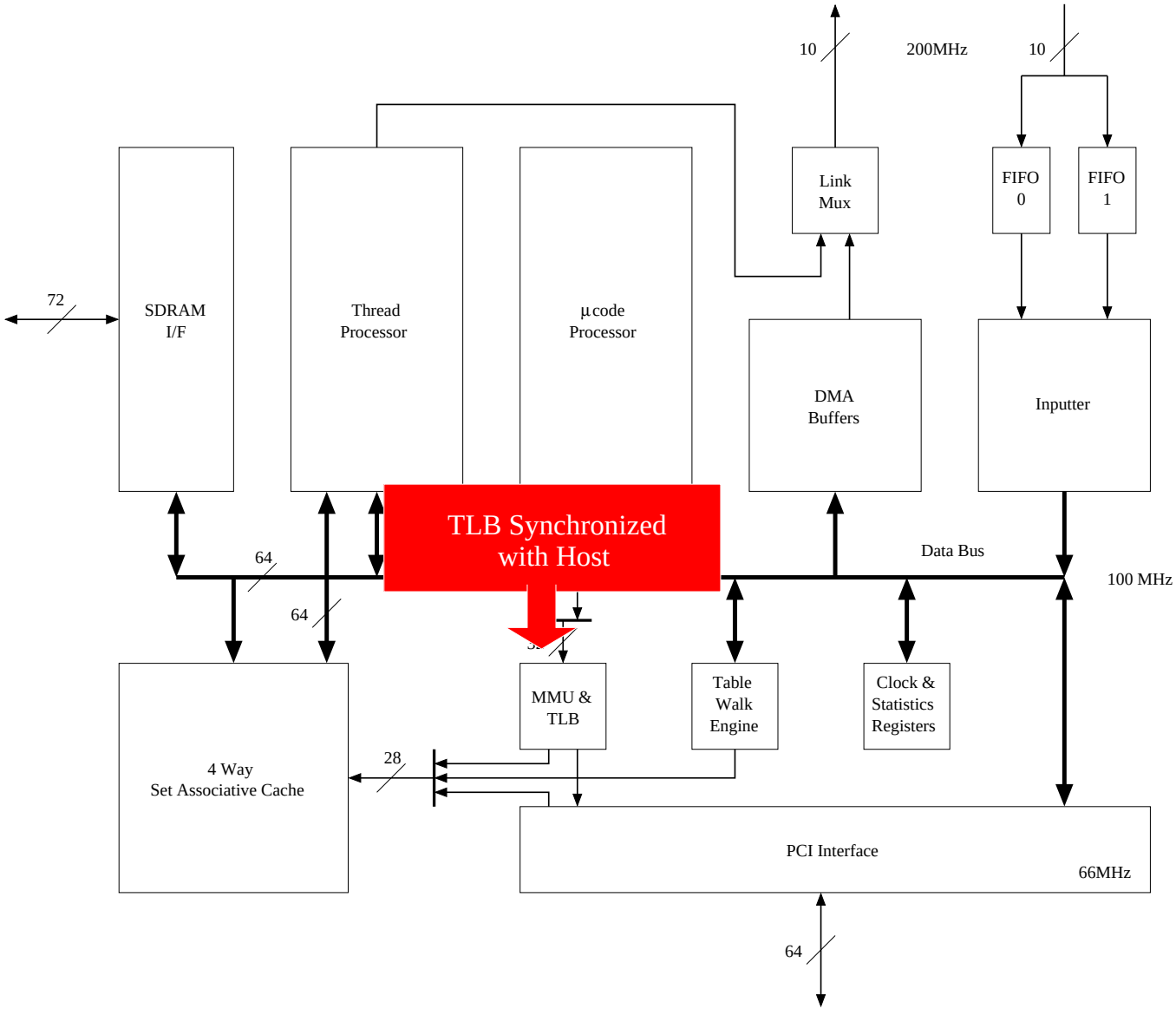
Quadrics Network Overview

- QsNET (Elan3) provides an abstraction of distributed virtual shared memory
- Each process can map a portion of its address space into the global memory
- These address spaces constitutes the virtual shared memory
- This shared memory is fully integrated with the native operating system
- Based on two building blocks:
  - a network interface card called **Elan**
  - a crossbar switch called **Elite**

Quadrics Network: Elan



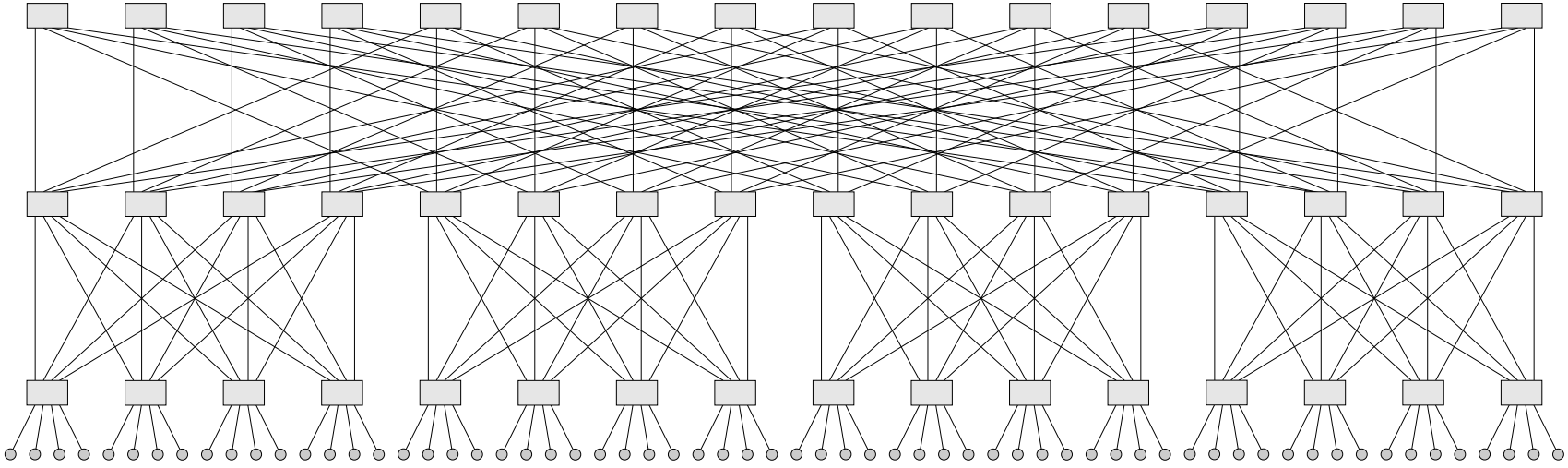
Quadrics Network: Elan



Quadrics Network: Elite

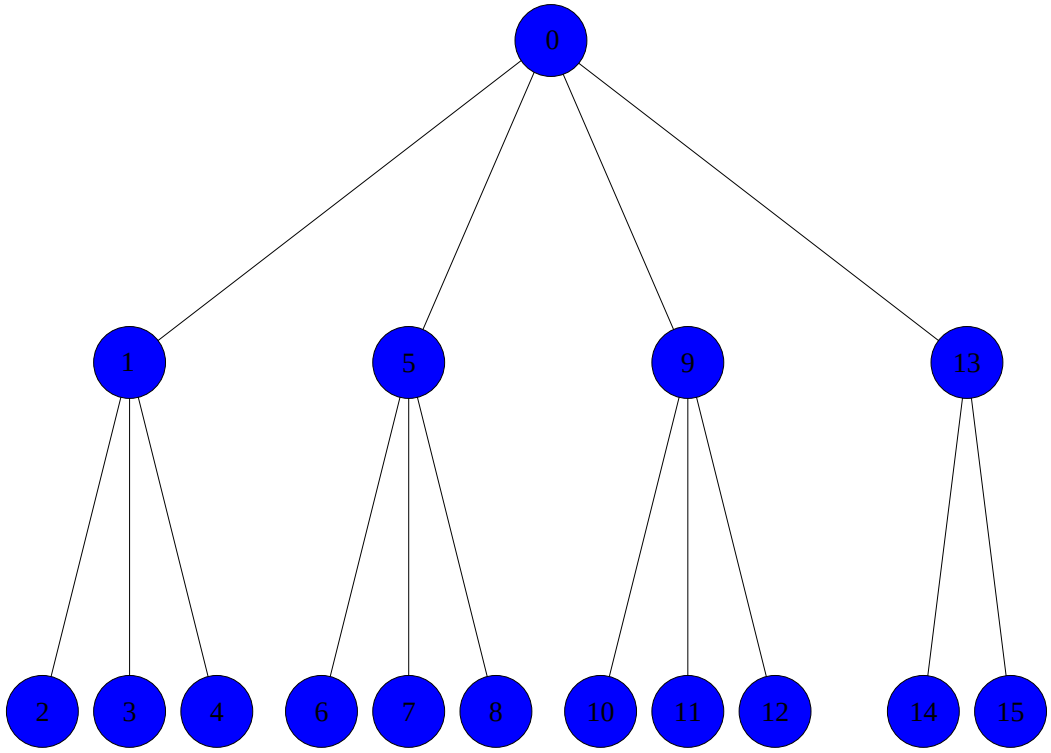
- 8 bidirectional links with 2 virtual channels in each direction
- An internal 16x8 full crossbar switch
- 400 MB/s on each link direction
- Packet error detection and recovery, with routing and data transactions CRC protected
- 2 priority levels plus an aging mechanism
- Adaptive routing
- Hardware support for broadcast

Quaternary fat-tree



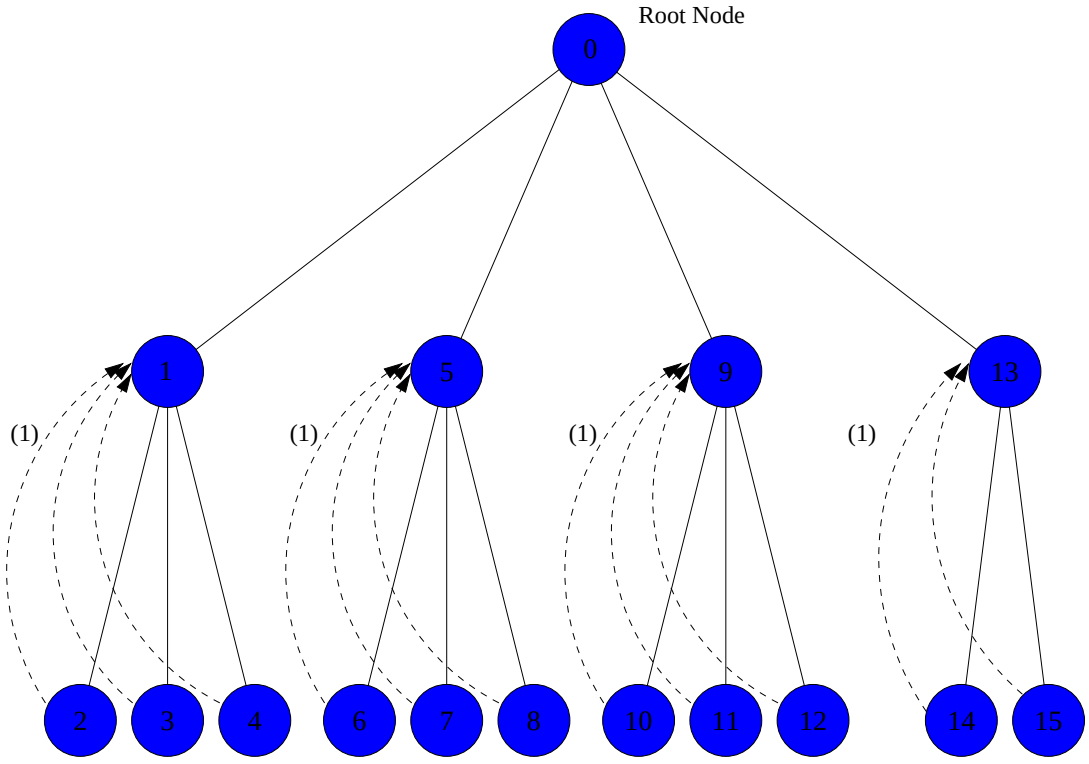
- Elans and Elites are connected in a fat-tree topology

Software-Based Barrier



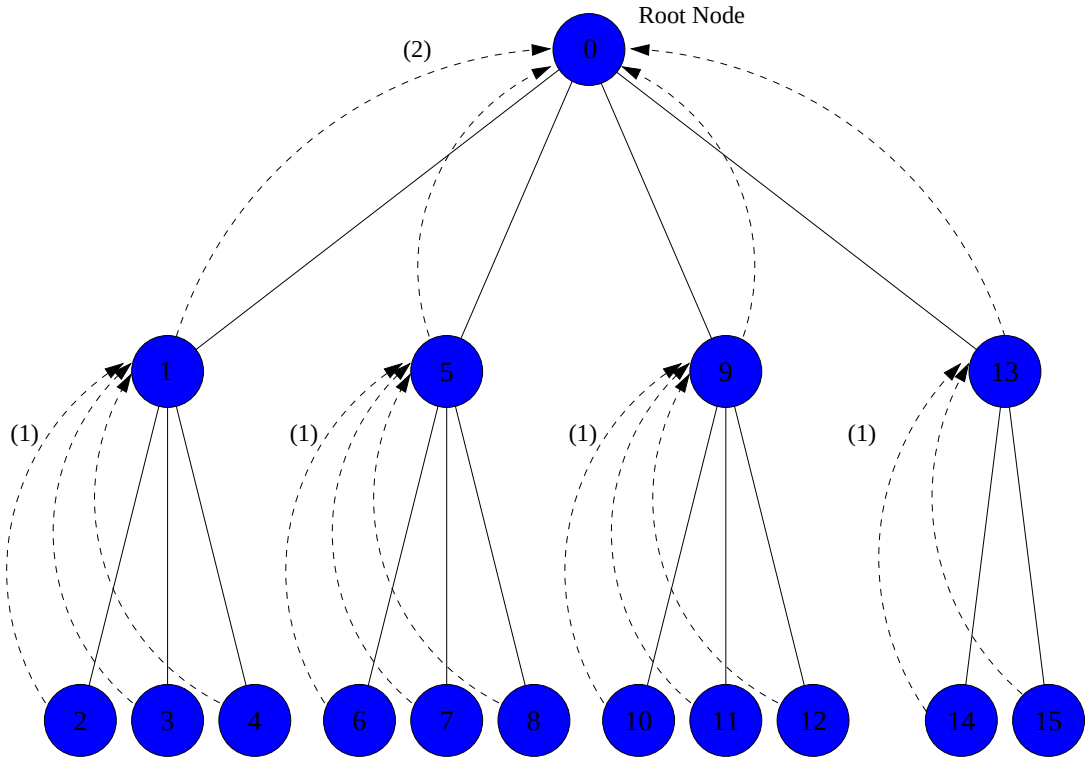
- The software-based barrier is executed is using point to point messages
- These messages are sent from Elan to Elan, without interrupting the processing node

Software-Based Barrier



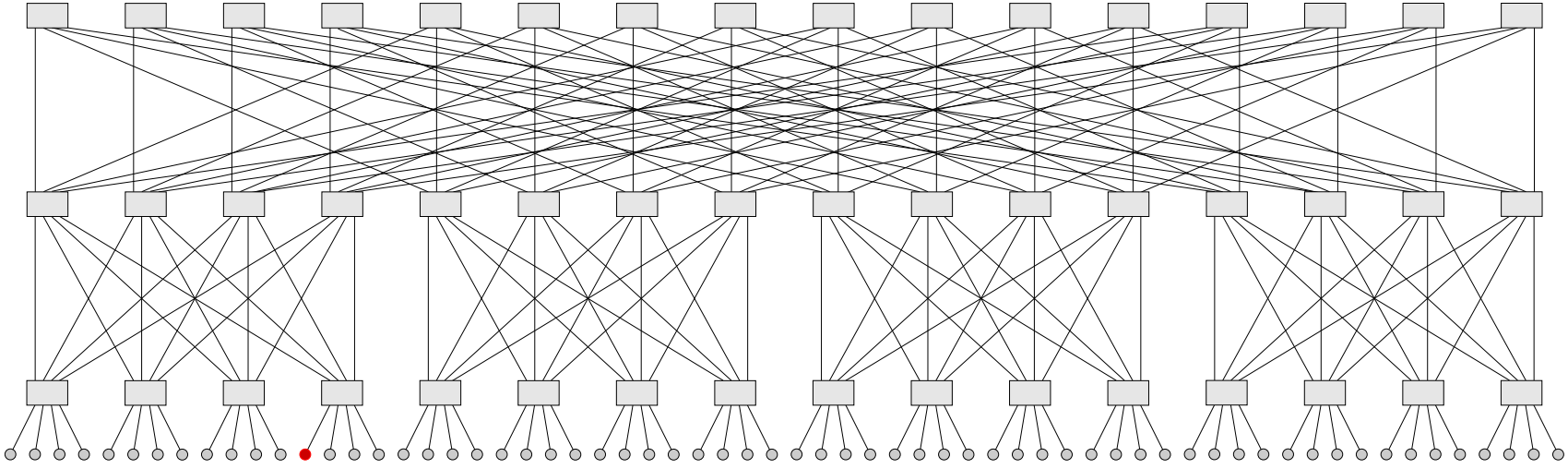
- Each Elan Network Interface waits for ’ready’ signals from its children (1) ...

Software-Based Barrier



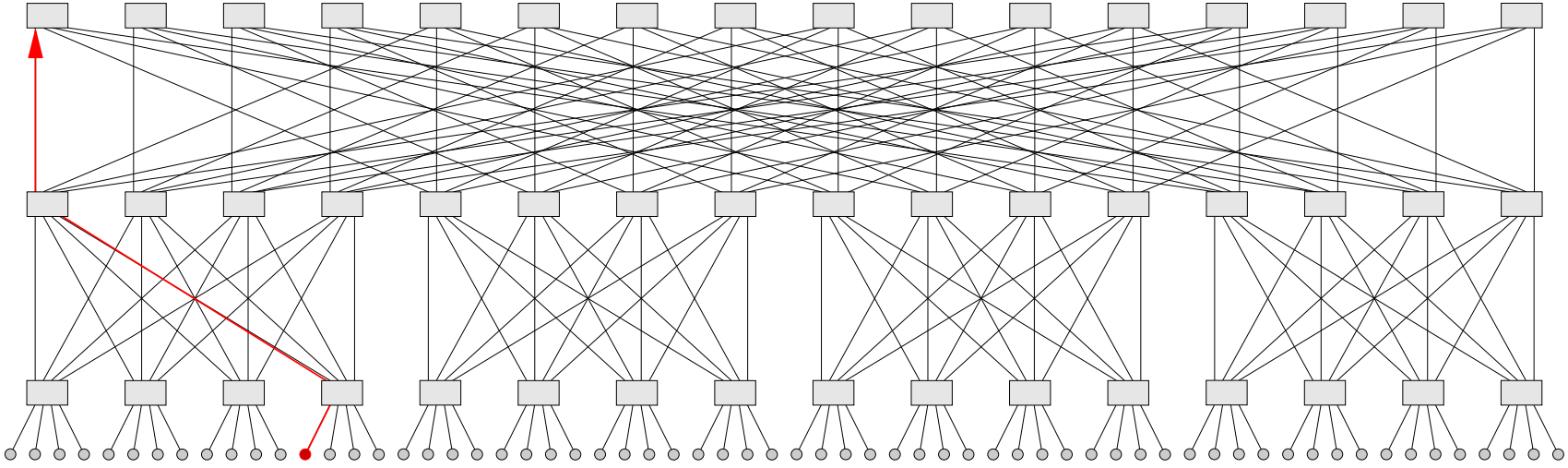
- ... and sends its own signal up to the parent process (2)

Hardware-Based Barrier



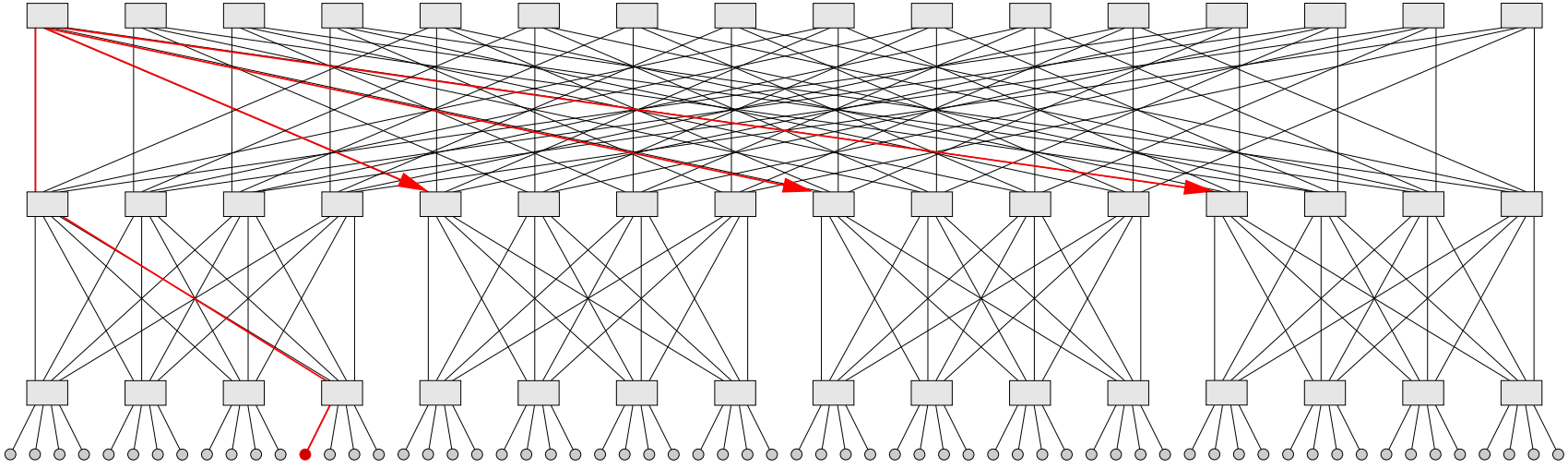
- The root node sends a multicast packet

Hardware-Based Barrier



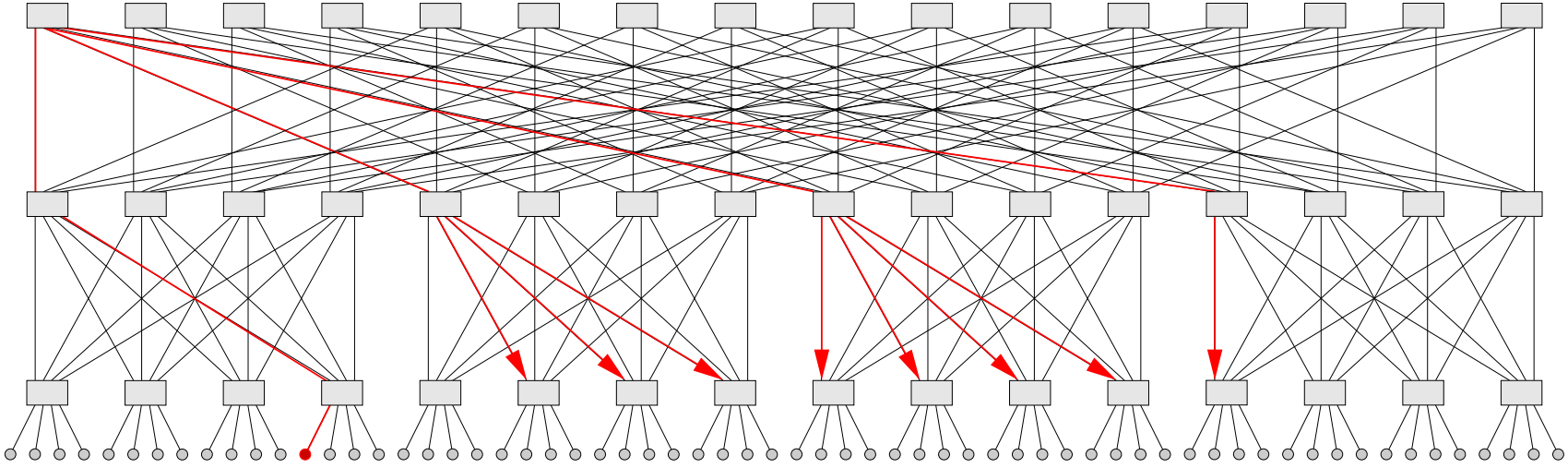
- The packet reaches the top of the tree

Hardware-Based Barrier



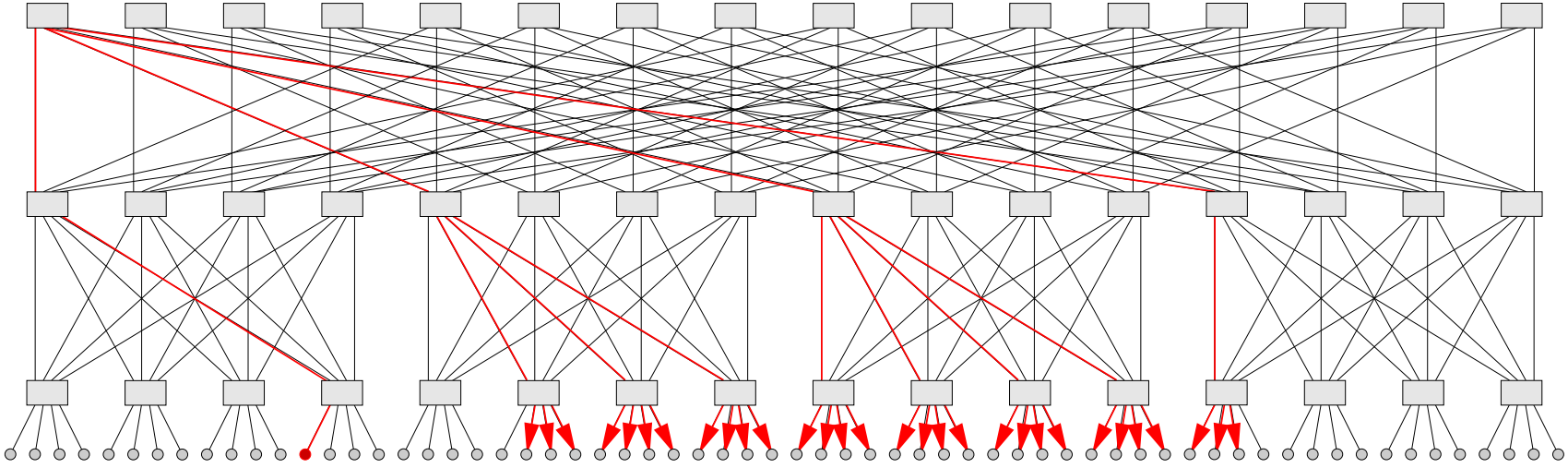
- The packet is multicast down the logical tree

Hardware-Based Barrier



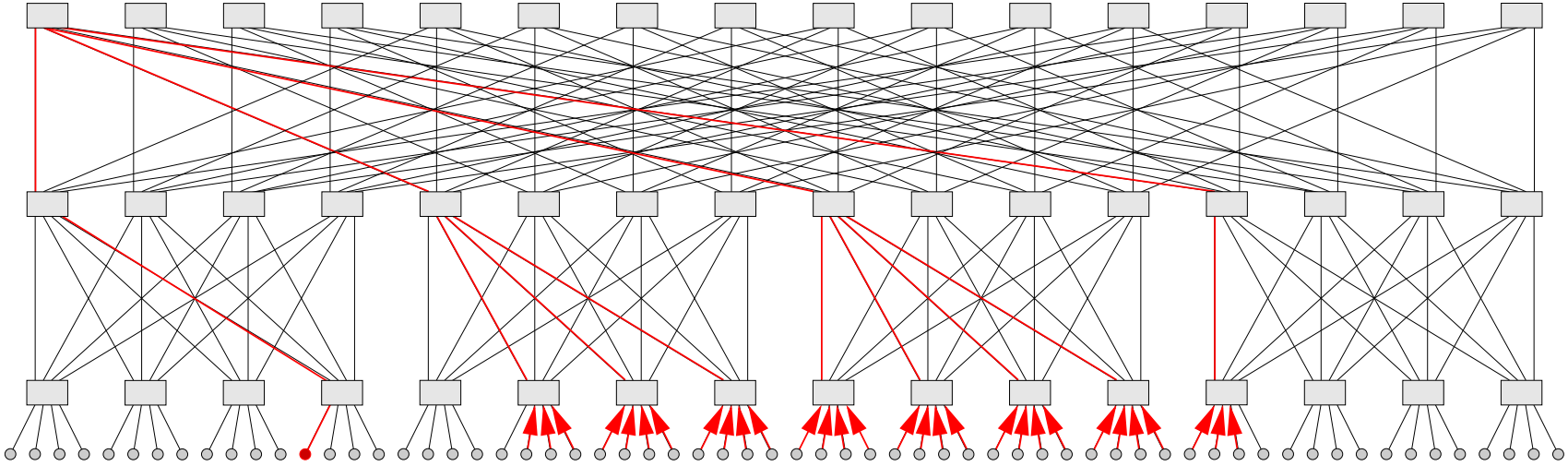
- The packet is multicast down the logical tree

Hardware-Based Barrier



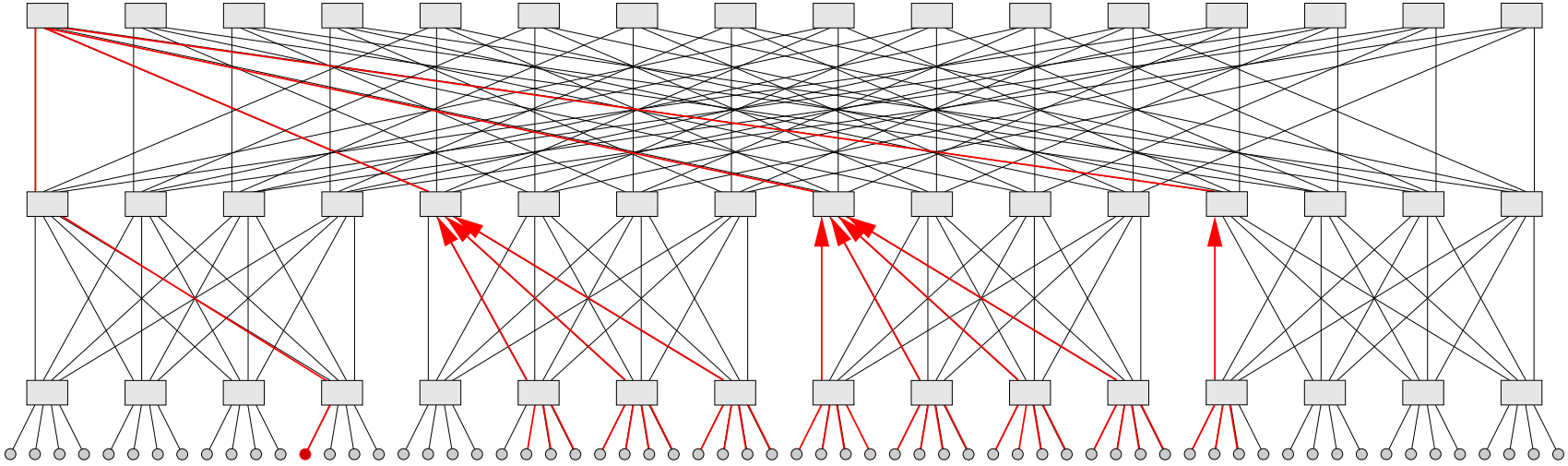
- The packet is multicast down the logical tree

Hardware-Based Barrier

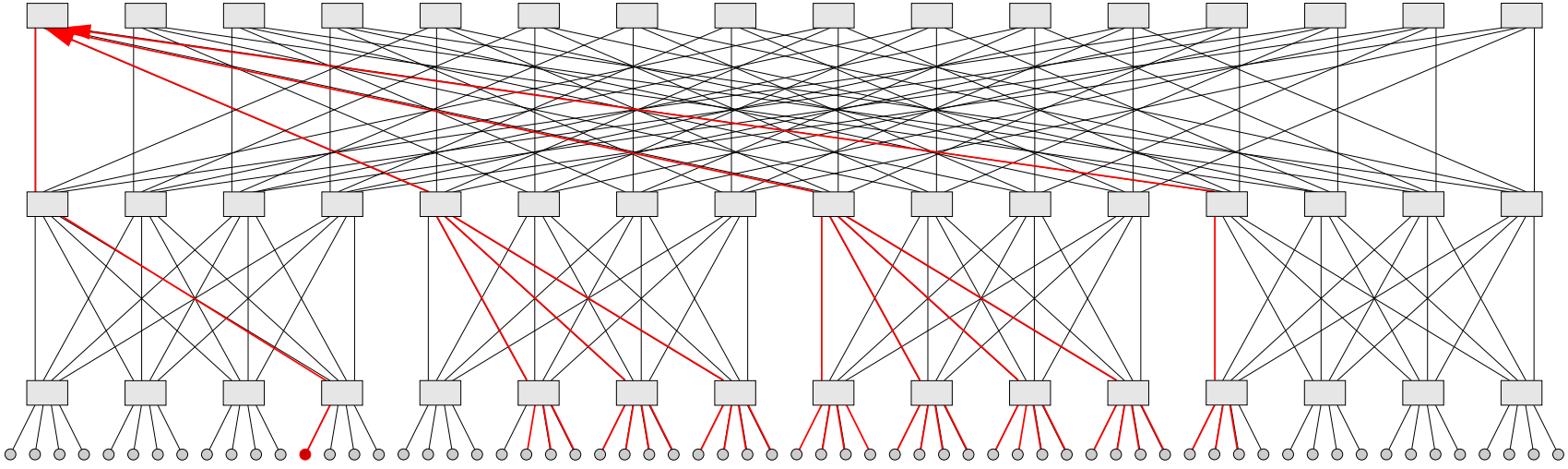


- The results of the collective operation are combined and sent back to the root.
- The tree of circuits is active during the whole collective communication.

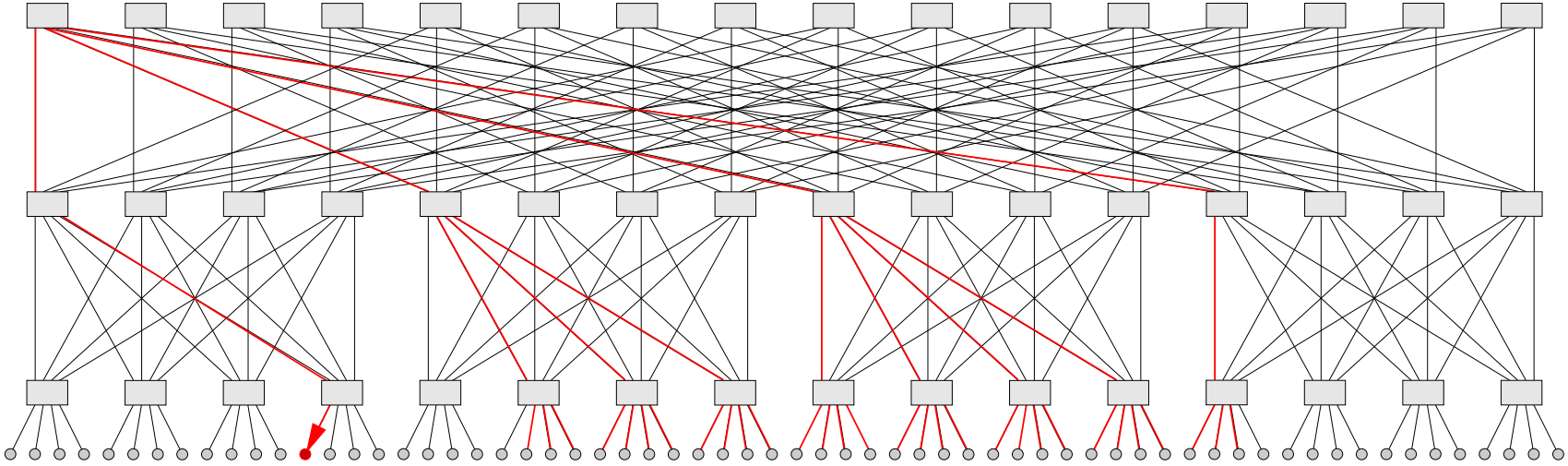
Hardware-Based Barrier



Hardware-Based Barrier



Hardware-Based Barrier

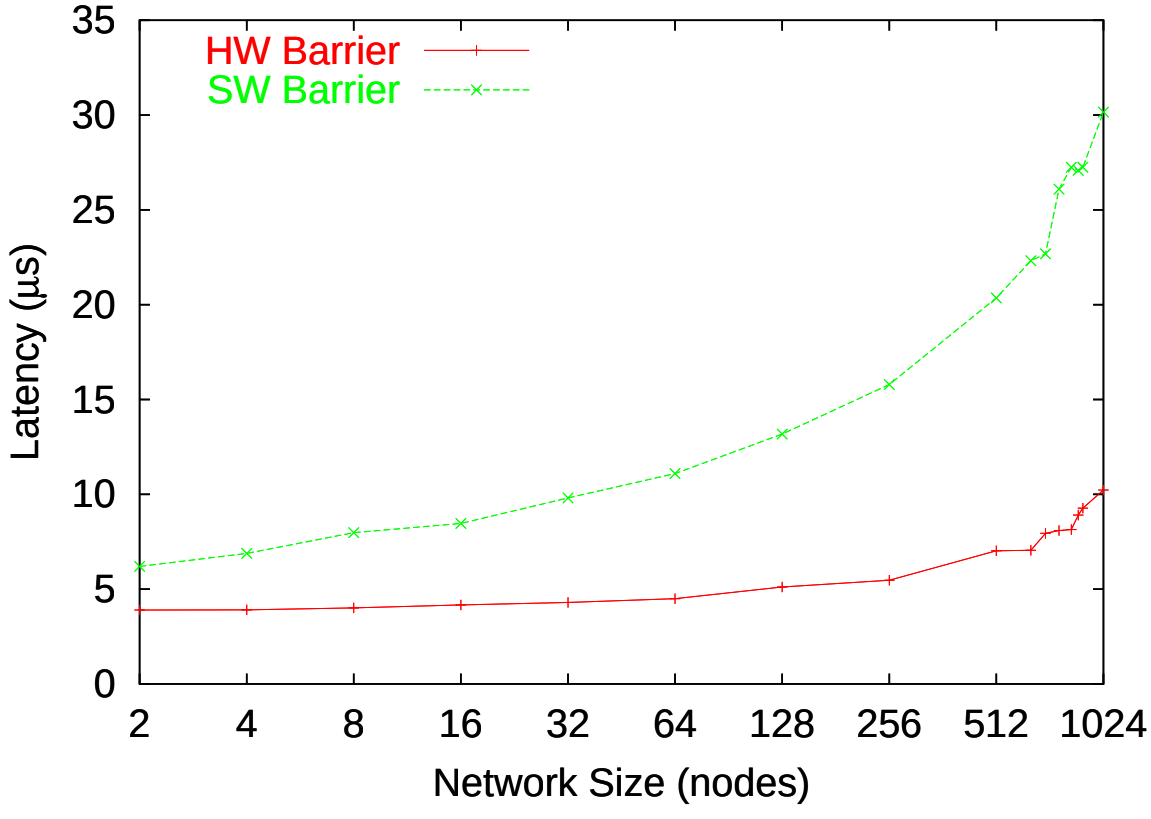


- The final result reaches the root
- The whole collective communication is atomic

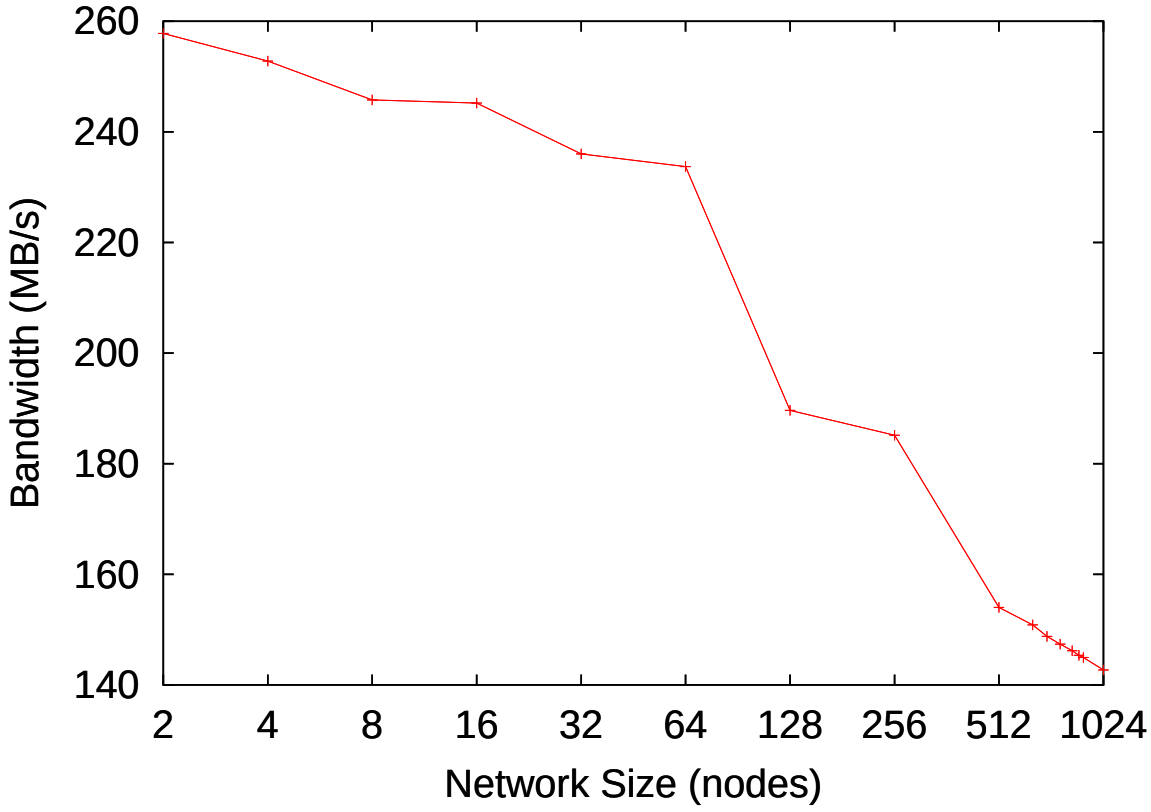
Performance and Scalability

- We report performance and scalability results of three common collective communication patterns on a 1024-node segment of the Q machine
  - Barrier Synchronization
  - Broadcast (one to all)
  - Hot-spot (all to one)

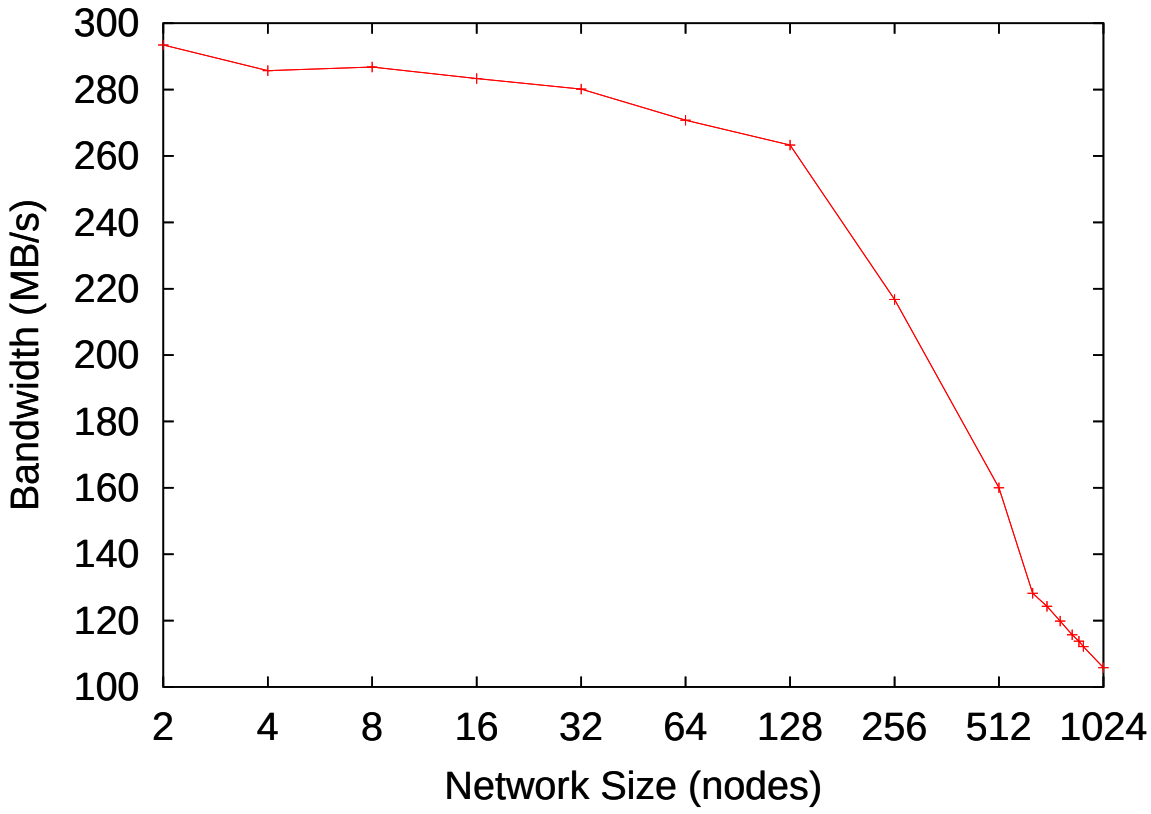
Barrier Synchronization



Broadcast


260
240
220
200
180
160
140
Bandwidth (MB/s)
2
4
8
16
32
64
128
256
512
1024
Network Size (nodes)

Hot Spot


300
280
260
240
220
200
180
160
140
120
100
Bandwidth (MB/s)
2
4
8
16
32
64
128
256
512
1024
Network Size (nodes)

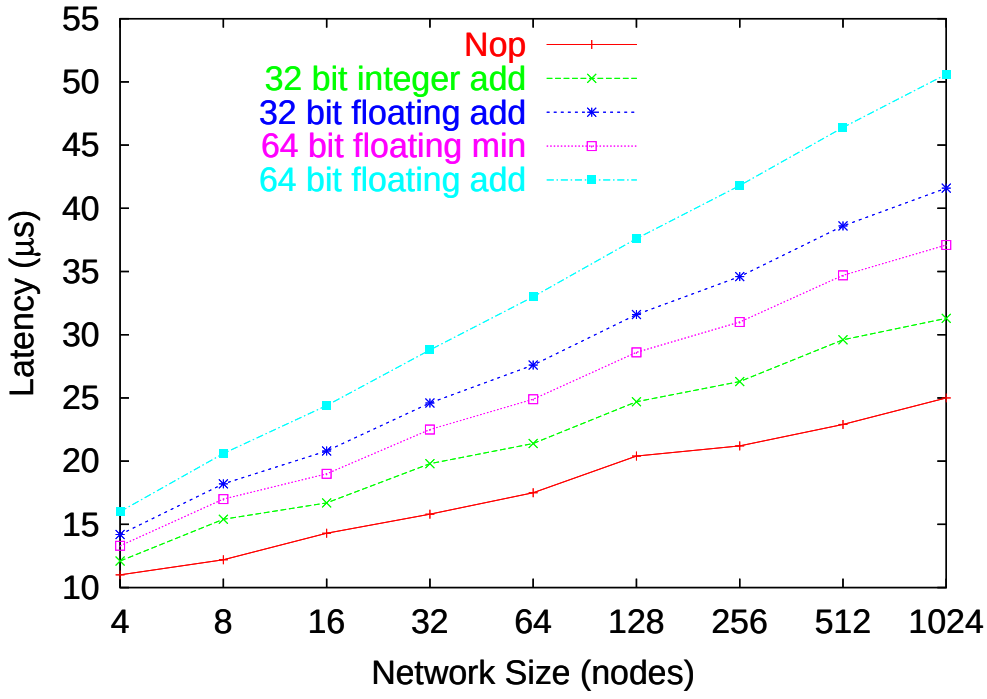
Allreduce

- The most common collective operation in many ASCI codes is *allreduce*
- Sage performs an allreduce every few milliseconds (the frequency is influenced by the input deck)
- The largest vector size is only 6 elements, but in general it uses vectors of only one element
- The performance and the scalability of the allreduce is very important for Sage.

Allreduce in the NIC?

- The processes in the host that perform the allreduce can be interrupted/descheduled by the OS. We try to alleviate this problem with a network-based solution.
- Optimize the common case
- We managed to implement IEEE compliant floating point in the Elan (fully compliant up to 64 bits, and partially up to 80 bits)
- We have implemented an allreduce in the NIC (which is not affected by the noise in the processing nodes, to a certain extent)
- expected manifold performance improvement in a large scale machine

Reduce



- The graph describes measured (up to 32 nodes) and expected performance of the NIC-based allreduce algorithm.
- The best performance obtained on 1024 nodes of the Q machine is about 300 μs.

Conclusions

- We presented an overview of both software- and hardware-based collective communication algorithms on the Quadrics network
- We also presented some scalability and performance results of three collective primitives, barrier, broadcast and hot spot on a 1024-node segment of the Q machine
- Finally, we discussed some preliminary results on the implementation of the *allreduce*, a common operation on many ASCI codes, in the network interface card.

Resources

More information can be found at the following URLs:

Quadrics network

http://www.quadrics.com

http://www.c3.lanl.gov/~fabrizio/publications.html

PAL publication page

http://www.c3.lanl.gov/par_arch